Caractérisation d'observables interprétables en vue de l'analyse stylométrique de corpus anciens

Anaëlle Baledent

Sorbonne Université

22 mai 2019

Plan

Contexte

Analyse Stylométrique

- Pour nous : **Outiller** l'expert
- Jeu d'indices stylistiques :
  - extraits d'une liste prédéfinie
    [Stamatatos, 2009, Lecluze and Lejeune, 2014]
  - calculés en fonction des données
    [Quiniou et al., 2012, Brixtel et al., 2015]
- Combinet efficaité et interprétabilité

Notre travail : étude du style de Dumas (père)

- Étudier ses caractéristiques . . .
- en le comparant à un des contemporains de Dumas (P.Féval)

Contexte

Analyse Stylométrique

- Pour nous : **Outiller** l'expert
- Jeu d'indices stylistiques :
  - extraits d'une liste prédéfinie
    [Stamatatos, 2009, Lecluze and Lejeune, 2014]
  - calculés en fonction des données
    [Quiniou et al., 2012, Brixtel et al., 2015]
- Combinet efficaité et interprétabilité

Notre travail : étude du style de Dumas (père)

- Étudier ses caractéristiques . . .
- en le comparant à un des contemporains de Dumas (P.Féval)

Comment représenter ces caractéristiques ?

Observables

Définition

- Des unités observables, détectables automatiquement, caractéristiques d'un style (**individuel** ou **collectif**)
- Différentes unités d'analyse : phonème, morphème. . .

Notre choix : utiliser des séquences syntaxiques

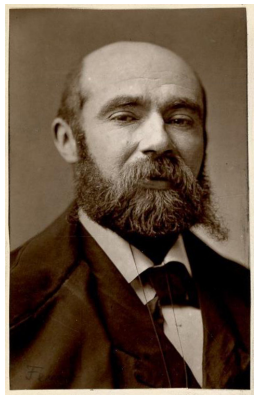
- Liste d'étiquettes morphosyntaxiques

	D'Artagnan	raconte	sa	première	visite	à	Mr	de	Tréville
→	NPP	V	DET	ADJ NC	P	NC	P	NPP	

- NPP-V-DET (effectif = 1, longueur = 3)
- NC-P (effectif = 2, longueur = 2)
- . . .

Plan

Le Corpus Dumas Féval (CDF)



Corpus : Dumas

- 9 œuvres
- Entre 1839-1850
- Historique, aventure
- 1 267 651 tokens

Contre-corpus : Féval

- 10 œuvres
- Entre 1842-1868
- Historique, aventure
- 1 236 781 tokens

Statistiques détaillées du CDF

Titre	Année	Genre	# tokens	# chapitres
DUMAS				
Le capitaine Paul	1838	Aventure	76 787	19
Acté	1839	Historique	81 972	19
Othon l'archer	1840	Historique, Aventure	**38 619**	11
Le chevalier d'Harmental	1843	Historique, Aventure	164 631	48
Les trois mousquetaires	1844	Historique, Aventure	**284 527**	68
La reine Margot	1845	Historique	**258 859**	56 [1]
Le chevalier de Maison-Rouge	1846	Historique	**201 815**	56
La tulipe noire	1850	Historique, Aventure	89 471	33
La femme au collier de velours	1851	Fantastique, Historique	69 970	17 [2]
FEVAL				
Le médecin bleu	1842	Historique	**16 556**	9
Les fanfarons du roi	1843	Historique, Aventure	97 080	25
Le loup blanc	1843	Historique, Aventure	92 780	24
La fée des grèves	1850	Historique	95 336	34
La reine des épées	1852	Aventure	113 056	23
Le bossu	1857	Historique, Aventure	**270 531**	62
Les errants de la nuit	1857	Aventure	121 054	30
Le roi des gueux	1859	Historique, Aventure	162 748	25
La vampire	1865	Fantastique	109 227	27
Le cavalier fortune	1868	Aventure	158 413	59

1. 2 volumes
2. nouvelles

Classification non supervisée

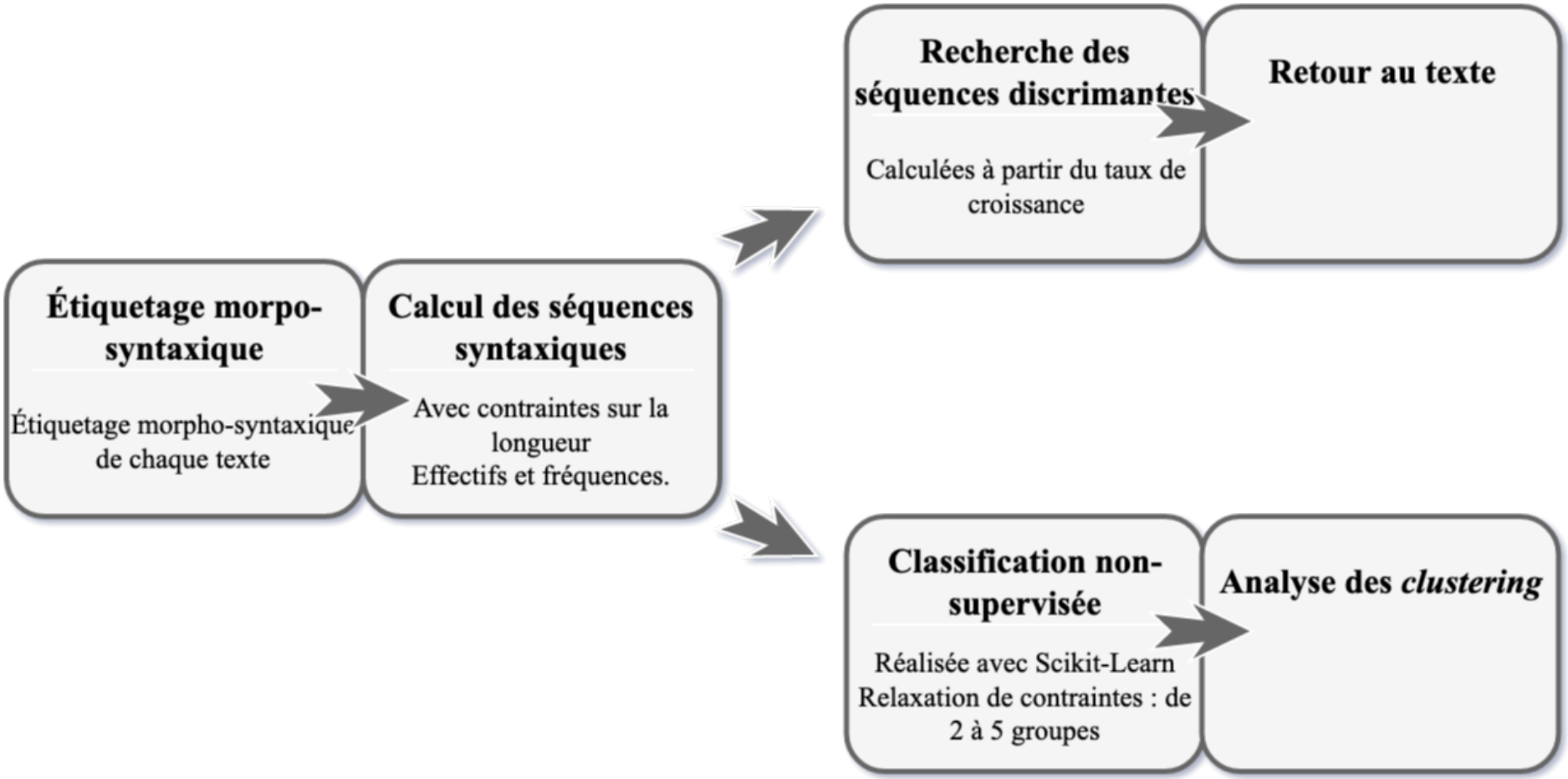
Classification . . .

- . . . supervisée : **classer** les données (thème, genre. . .)
- . . . non supervisée (*clustering*) : **regrouper** les données sans chercher une étiquette pré-définie

Intérêt de la non-supervision

- **Laisser parler les données**
- Ressortir des propriétés attendues et en découvrir de nouvelles
- Permettre l'analyse de corpus non étiquetés

Chaîne de traitement



Résultats

Expériences réalisées

- Exploitation des séquences syntaxiques dans le CDF
  - Séquences discriminantes et retour au texte
  - Classification non-supervisée
  - Visualisation : analyse en composantes principales
  - Visualisation : classification ascendante hiérarchique
- Études des textes du CDF à différents grains
  - *Clustering* des chapitres introductifs et conclusifs
  - Explorer la séquentialité d'une œuvre (*Les trois mousquetaires*)
  - Retrouver des parties définies (*Le bossu*)
- Clustering Ascendant Hiérarchique des premiers/derniers chapitres selon le lexique
  - Delta de Burrows
  - Delta cosinus

`https://github.com/AnaelleBaledent/analyse-stylometrique_dumas-feval`

Expériences réalisées

- Exploitation des séquences syntaxiques dans le CDF
  - **Séquences discriminantes et retour au texte**
  - **Classification non-supervisée**
  - **Visualisation : analyse en composantes principales**
  - Visualisation : classification ascendante hiérarchique
- Études des textes du CDF à différents grains
  - *Clustering* des chapitres introductifs et conclusifs
  - Explorer la séquentialité d'une œuvre (*Les trois mousquetaires*)
  - **Retrouver des parties définies (*Le bossu*)**
- Clustering Ascendant Hiérarchique des premiers/derniers chapitres selon le lexique
  - Delta de Burrows
  - Delta cosinus

`https://github.com/AnaelleBaledent/analyse-stylometrique_dumas-feval`

Séquences discriminantes : exemples

Séquence propre à Dumas

- **NC-P-PROREL-CLS-V** : Nom commun - Préposition - Pronom relatif - Pronom clitique - Verbe

car j'aurai un **frère pour lequel je n'aurai** plus d'amour, et un **mari pour lequel je n'aurai** plus d'estime ! (*Le capitaine Paul*)

Séquence propre à Féval

- **PUNC-V-NPP-PROREL** : Ponctuation - Verbe - Nom propre - Pronom relatif

– Quoi ! sauvés tous deux ! sauvés par vous **! dit Sainte, qui** fondit en larmes. Que faire pour vous prouver ma reconnaissance ?
– Voulez-vous me rendre bien content **? dit Brand, qui** se sentit rougir sous le cuir bronzé de sa joue. (*Le médecin bleu*)

Clustering selon la fréquence des séquences

Longueurs des séquences :	min=4 & max=5			
Nombre de *clusters* :	2	3	4	5
Dumas/le-capitaine-paul	V	V	V	V
Dumas/acte	V	V	V	V
Dumas/othon-l-archer	V	V	V	Z
Dumas/le-chevalier-d-harmental	V	V	Y	Y
Dumas/les-trois-mousquetaires	V	V	Y	Y
Dumas/la-reine-margot	V	V	Y	Y
Dumas/le-chevalier-de-maison-rouge	V	V	Y	Y
Dumas/la-tulipe-noire	V	V	V	V
Dumas/la-femme-au-collier-de-velours	V	V	V	V
Feval/le-medecin-bleu	W	X	X	X
Feval/les-fanfarons-du-roi	W	W	W	W
Feval/le-loup-blanc	W	W	W	W
Feval/la-fee-des-greves	W	W	W	W
Feval/la-reine-des-epees	W	W	W	W
Feval/le-bossu	W	W	Y	Y
Feval/les-errants-de-nuit	W	W	W	W
Feval/le-roi-des-gueux	W	W	W	W
Feval/la-vampire	W	W	W	W
Feval/le-cavalier-fortune	W	W	W	W

Communication orale à Phraseorom (Erlangen, 2019)

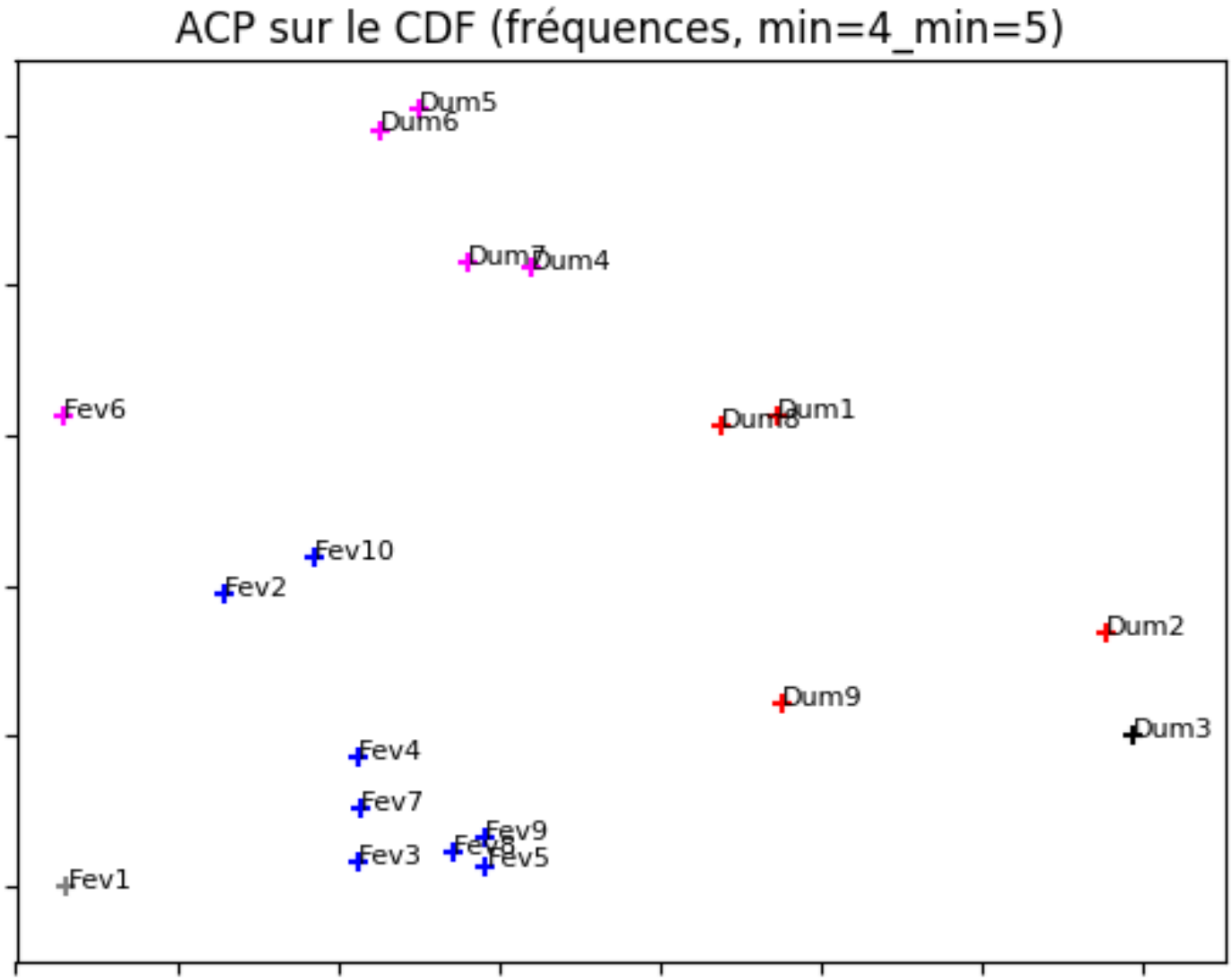
Clustering selon la fréquence des séquences

Longueurs des séquences :	min=4 & max=5			
Nombre de *clusters* :	2	3	4	5
Dumas/le-capitaine-paul	V	V	V	V
Dumas/acte	V	V	V	V
Dumas/othon-l-archer	V	V	V	Z
Dumas/le-chevalier-d-harmental	V	V	Y	Y
Dumas/les-trois-mousquetaires	V	V	Y	Y
Dumas/la-reine-margot	V	V	Y	Y
Dumas/le-chevalier-de-maison-rouge	V	V	Y	Y
Dumas/la-tulipe-noire	V	V	V	V
Dumas/la-femme-au-collier-de-velours	V	V	V	V
Feval/le-medecin-bleu	W	X	X	X
Feval/les-fanfarons-du-roi	W	W	W	W
Feval/le-loup-blanc	W	W	W	W
Feval/la-fee-des-greves	W	W	W	W
Feval/la-reine-des-epees	W	W	W	W
Feval/le-bossu	W	W	Y	Y
Feval/les-errants-de-nuit	W	W	W	W
Feval/le-roi-des-gueux	W	W	W	W
Feval/la-vampire	W	W	W	W
Feval/le-cavalier-fortune	W	W	W	W

Communication orale à Phraseorom (Erlangen, 2019)

Mieux appréhender la distance des œuvres et des *clusters* ?

Visualisation : Analyse en Composantes Principales (ACP)



Retrouver les six parties du *Bossu*

Partie 1		Partie 2		Partie 3		Partie 4		Partie 5		Partie 6	
Chapitre 1	C	Chapitre 9	C	Chapitre 20	C	Chapitre 30	C	Chapitre 40	E	Chapitre 53	D
Chapitre 2	C	Chapitre 10	E	Chapitre 21	F	Chapitre 31	E	Chapitre 41	C	Chapitre 54	E
Chapitre 3	E	Chapitre 11	E	Chapitre 22	E	Chapitre 32	E	Chapitre 42	B	Chapitre 55	E
Chapitre 4	E	Chapitre 12	E	Chapitre 23	F	Chapitre 33	E	Chapitre 43	E	Chapitre 56	E
Chapitre 5	E	Chapitre 13	E	Chapitre 24	F	Chapitre 34	E	Chapitre 44	F	Chapitre 57	F
Chapitre 6	B	Chapitre 14	F	Chapitre 25	F	Chapitre 35	C	Chapitre 45	A	Chapitre 58	E
Chapitre 7	E	Chapitre 15	E	Chapitre 26	F	Chapitre 36	D	Chapitre 46	B	Chapitre 59	D
Chapitre 8	E	Chapitre 16	E	Chapitre 27	F	Chapitre 37	F	Chapitre 47	D	Chapitre 60	D
		Chapitre 17	E	Chapitre 28	E	Chapitre 38	E	Chapitre 48	E	Chapitre 61	C
		Chapitre 18	E	Chapitre 29	E	Chapitre 39	E	Chapitre 49	E	Chapitre 62	C
		Chapitre 19	E					Chapitre 50	E		
								Chapitre 51	B		
								Chapitre 52	C		

Figure – Résultats du *clustering* selon la méthode des K-means sur les séquences ($4 <= len(sequence) <= 5$) des chapitres du *Bossu* (Féval). Lettres A à F : noms des clusters.

Retrouver les six parties du *Bossu*

Partie 1		Partie 2		Partie 3		Partie 4		Partie 5		Partie 6	
Chapitre 1	C	Chapitre 9	C	Chapitre 20	C	Chapitre 30	C	Chapitre 40	E	Chapitre 53	D
Chapitre 2	C	Chapitre 10	E	Chapitre 21	F	Chapitre 31	E	Chapitre 41	C	Chapitre 54	E
Chapitre 3	E	Chapitre 11	E	Chapitre 22	E	Chapitre 32	E	Chapitre 42	B	Chapitre 55	E
Chapitre 4	E	Chapitre 12	E	Chapitre 23	F	Chapitre 33	E	Chapitre 43	E	Chapitre 56	E
Chapitre 5	E	Chapitre 13	E	Chapitre 24	F	Chapitre 34	E	Chapitre 44	F	Chapitre 57	F
Chapitre 6	B	Chapitre 14	F	Chapitre 25	F	Chapitre 35	C	Chapitre 45	A	Chapitre 58	E
Chapitre 7	E	Chapitre 15	E	Chapitre 26	F	Chapitre 36	D	Chapitre 46	B	Chapitre 59	D
Chapitre 8	E	Chapitre 16	E	Chapitre 27	F	Chapitre 37	F	Chapitre 47	D	Chapitre 60	D
		Chapitre 17	E	Chapitre 28	E	Chapitre 38	E	Chapitre 48	E	Chapitre 61	C
		Chapitre 18	E	Chapitre 29	E	Chapitre 39	E	Chapitre 49	E	Chapitre 62	C
		Chapitre 19	E					Chapitre 50	E		
								Chapitre 51	B		
								Chapitre 52	C		

Figure – Résultats du *clustering* selon la méthode des K-means sur les séquences ($4 <= len(sequence) <= 5$) des chapitres du *Bossu* (Féval). Lettres A à F : noms des clusters.

→ Attirer l'attention sur des passages ou chapitres en particulier

Retour au texte

- Séquence syntaxique : **NP_V**
- Extrait (*Les trois mousquetaires*, épilogue) :

Bazin	***devint***	*frère lai.*
Athos	***resta***	*mousquetaire [. . .]*
Grimaud	***suivit***	*Athos.*
D'Artagnan	***se battit***	*trois fois avec Rochefort et le blessa trois fois.*
Planchet	***obtint***	*de Rochefort le grade de sergent dans les gardes.*
M. Bonacieux	***vivait***	*fort tranquille [. . .]*

- Application : analyse des séries homogènes de cadres du discours ([Vigier et al., 2009], [Charolles, 2009])

Conclusions et perspectives

Contributions principales de mon mémoire

- *Clustering* fondé sur les séquences syntaxiques : meilleure généralisation que le lexique
- Incidence de la longueur des textes : regroupements liés à la variété (richesse ?) syntaxique
- Changement de granularité : attire l'attention de l'expert sur certains passages
- Meilleure visualisation des *clusters* : ACP

Code `https://github.com/AnaelleBaledent/analyse-stylometrique_dumas-feval`

Contributions principales de mon mémoire

- *Clustering* fondé sur les séquences syntaxiques : meilleure généralisation que le lexique
- Incidence de la longueur des textes : regroupements liés à la variété (richesse ?) syntaxique
- Changement de granularité : attire l'attention de l'expert sur certains passages
- Meilleure visualisation des *clusters* : ACP

Code `https://github.com/AnaelleBaledent/analyse-stylometrique_dumas-feval`

Question de recherche

Reproduire l'expérience sur les Mazarinades (1648-1653) ?

Perspectives

L'importance de l'état du corpus

- Accès au mode texte ?
- État des documents et de la langue
- Corpus non bruité VS corpus bruité

Des observables adaptables aux données traitées

- Se concentrer sur les données telles qu'elles sont disponibles (*corpus integrity* [Dias, 2010])
- Déterminer la chaîne de traitement optimale selon l'état du corpus
- Adapter les observables aux corpus

Réalisations

- **Analyse stylistique automatique : à la recherche d'indices efficaces et pertinents pour caractériser le style de Dumas** Anaëlle Baledent et Gaël Lejeune, Actes de Phraseorom 2019, à paraître
- Code disponible en ligne

  `https://github.com/AnaelleBaledent/analyse-stylometrique_dumas-feval`

Merci de votre attention

Bibliographie I

Brixtel, R., Lecluze, C., and Lejeune, G. (2015).
Attribution d'Auteur : approche multilingue fondée sur les répétitions maximales.
In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*, pages 208–219.

Charolles, M. (2009).
Les cadres de discours comme marques d'organisation des discours.
In Venier, F., editor, *Tra Pragmatica e Linguistica Testuale*, pages 401–409. Edizioni dell'Orso, Alessandria.

Dias, G. (2010).
Information Digestion.
Habilitation à diriger des recherches en mathématiques et en informatique, Université d'Orléans.

Lecluze, C. and Lejeune, G. (2014).
Deft 2014, analyse automatique de textes littéraires et scientifiques en langue française.
In *Actes de DEFT 2014 : 10ème DÉfi Fouille de Textes*, pages 11–19, Marseille, France.

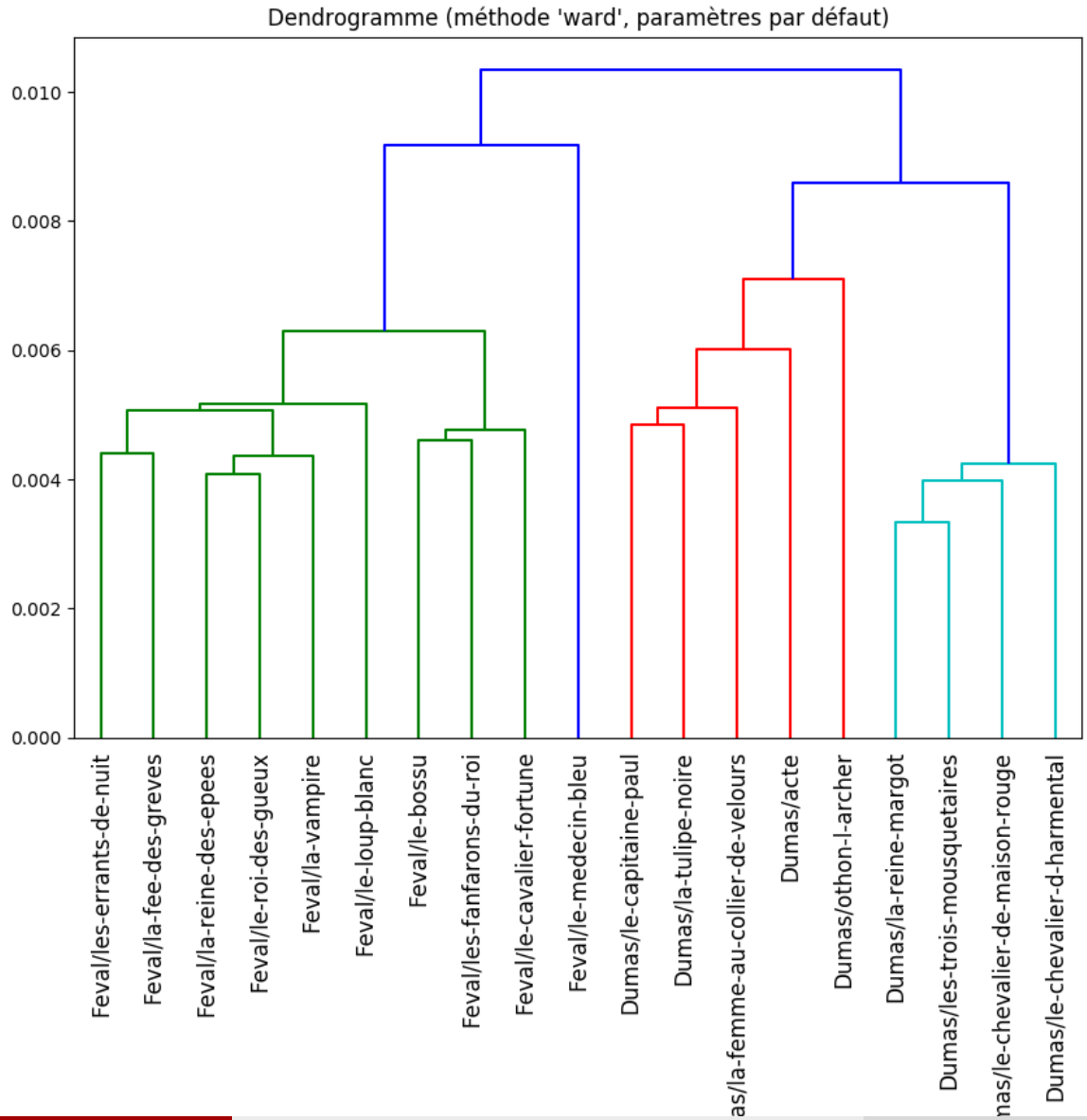
Bibliographie II

Quiniou, S., Cellier, P., Charnois, T., and Legallois, D. (2012).
What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics ?
In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'12)*, pages 166–177, New Delhi, India.

Stamatatos, E. (2009).
A survey of modern authorship attribution methods.
Journal of the American Society for Information Science and Technology, 60(3) :538–556.

Vigier, D., Ferrari, S., and Charnois, T. (2009).
Ce que le texte fait aux cadres de discours : le cas des séries homogènes de cadres.
In *Communication aux Journées Scientifiques du CRISCO organisées par D. Legallois & F. Neveu*, Caen, France. D. Legallois & F. Neveu.

Visualisation : dendrogramme (méthode Ward)



CAH des premiers/derniers chapitres avec Stylo (R)